s.16(2)
s.21(1)(b)

2019 MSBU 001266
UNCLASSIFIED
IRBV

MEMORANDUM TO THE DEPUTY MINISTER OF EMPLOYMENT AND SOCIAL DEVELOPMENT

Possibility of Using Machine Learning and Artificial Intelligence to Identify an Individual in an Anonymized Dataset

Possibilité d'utiliser l'apprentissage automatique et l'intelligence artificielle afin d'identifier un individu dans un jeu de données anonymisées

(FOR INFORMATION)

SUMMARY

Following the release of an MIT Technology Review article on the use of Machine Learning and Artificial Intelligence to reverse-engineer anonymization in datasets, researchers at the Université catholique de Louvain and the Imperial College London (Annex A) employed Machine Learning to re-identify individuals from a series of linked and de-identified datasets. The paper demonstrates that by using several linked datasets, a U.S. citizen could be correctly located in a de-identified database 81% of the time.

.../2

Canada

BACKGROUND

On July 23, 2019, the journal of *Nature Communications*, as cited in the MIT Technology Review (Annex A) published a research paper that states that researchers from Imperial College London and the *Université catholique de Louvain* have successfully created a machine-learning model that can be used to re-identify individuals from a de-identified dataset. De-identification refers to a spectrum of data methods (e.g., removing birthdates and SIN numbers from data) to remove personal identifiers of an individual in the data. The tool was created by assembling a database of 210 different datasets from five sources, including the U.S. Census. The datasets used in the study were de-identified (personal identifiers were removed) however, included personal information such as birthdate, zip code and gender. On average, using the provided datasets, the paper concluded that a U.S. citizen could be correctly located in a de-identified database 81% of the time. On a more granular level, given 15 demographic attributes of someone living in Massachusetts, there was a 99.98% chance of positive re-identification.

Artificial Intelligence (AI), as well as other techniques, are tools that can be used to attempt the identification of individuals in de-identified datasets, meaning datasets where personal identifiers have been stripped from the data. While increasingly sophisticated techniques are being applied that could increase the risk of re-identification in an anonymized dataset,

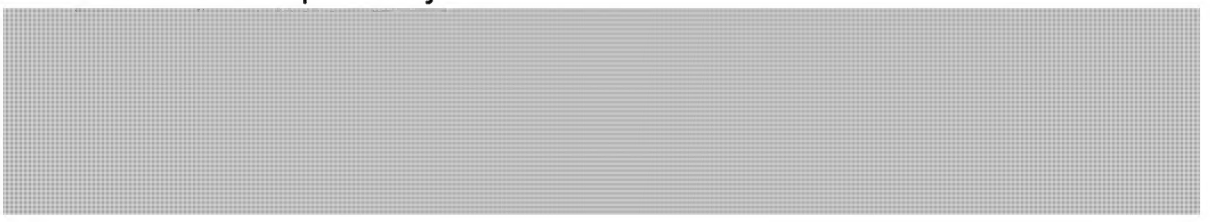
CURRENT STATUS

Two key factors are considered when determining the risk associated with re-identification. The first factor is the content of the data itself (i.e., the level of personal identifiable information), and the techniques used to anonymize this information. Anonymization refers to data methods that go beyond de-identification (e.g. masking, aggregation) to ensure no serious possibility of re-identifying an individual is possible.

The second factor is the risk brought about by the controls and environment where the data is being used, accessed or shared. This may include factors such as who is accessing the data, and the environment in which they are working. A few examples of controls that can be used to mitigate this contextual risk are training and skills development, levels of security clearance, and monitoring and logging activities, to name a few.

.../3

- The environment in which users get to interact with data is also an important component of risk mitigation, and greatly determines whether complex attack scenarios are a possibility.



CONSIDERATIONS

[redacted]

Although AI and Machine Learning are methods that can be used to identify individuals in anonymized datasets, the underlying issue is about re-identifying individuals, independent of the technologies or techniques applied to do so. Where data is required, a variety of checks and controls are used to mitigate re-identification risk. [redacted]

[redacted]

The CDO has also evaluated Statistics Canada's approach [redacted] and has consulted with senior methodologists at Statistics Canada. ESDC is in agreement with their approach to data access [redacted]

CONCLUSION

In conclusion, while increasingly sophisticated tools are being employed to re-identify individuals in data, ESDC is taking appropriate steps [redacted]

.../5

[redacted]

Senior Assistant Deputy Minister
Strategic and Service Policy Branch

Key Contact: Doug Woodworth
1 (506) 247-0629

Attachment: 1

You're very easy to track down, even when your data has been anonymized

A new study shows you can be easily re-identified from almost any database, even when your personal details have been stripped out.

by **Charlotte Jee**
Jul 23, 2019

https://www.technologyreview.com/s/613996/youre-very-easy-to-track-down-even-when-your-data-has-been-anonymized/

The data trail we leave behind us grows all the time. Most of it isn't that interesting—the takeout meal you ordered, that shower head you bought online—but some of it is deeply personal: your medical diagnoses, your sexual orientation, or your tax records.

The most common way public agencies protect our identities is anonymization. This involves stripping out obviously identifiable things such as names, phone numbers, email addresses, and so on. Data sets are also altered to be less precise, columns in spreadsheets are removed, and "noise" is introduced to the data. Privacy policies reassure us that this means there's no risk we could be tracked down in the database.

However, a new study in Nature Communications suggests this is far from the case.

Researchers from Imperial College London and the University of Louvain have created a machine-learning model that estimates exactly how easy individuals are to reidentify from an anonymized data set. You can check your own score here, by entering your zip code, gender, and date of birth.

On average, in the US, using those three records, you could be correctly located in an "anonymized" database 81% of the time. Given 15 demographic attributes of someone living in Massachusetts, there's a 99.98% chance you could find that person in any anonymized database.

"As the information piles up, the chances it isn't you decrease very quickly," says Yves-Alexandre de Montjoye, a researcher at Imperial College London and one of the study's authors.

The tool was created by assembling a database of 210 different data sets from five sources, including the US Census. The researchers fed this data into a machine-learning model, which

learned which combinations are more nearly unique and which are less so, and then assigns the probability of correct identification.

This isn't the first study to show how easy it is to track down individuals from anonymized databases. A paper back in 2007 showed that just a few movie ratings on Netflix can identify a person as easily as a Social Security number, for example. However, it shows just how far current anonymization practices have fallen behind our ability to break them. The fact that the data set is incomplete does not protect people's privacy, says de Montjoye.

It isn't all bad news. These same reidentification techniques were used by journalists working at the New York Times earlier this year to expose Donald Trump's tax returns from 1985 to 1994. However, the same method could be used by someone looking to commit ID fraud or obtain information for blackmail purposes.

"The issue is that we think when data has been anonymized it's safe. Organizations and companies tell us it's safe, and this proves it is not," says de Montjoye.

For peace of mind, companies should be using differential privacy, a complex mathematical model that lets organizations share aggregate data about user habits while protecting an individual's identity, argues Charlie Cabot, research lead at the privacy engineering firm Privitar.

The technique will get its first major test next year: it's being used to secure the US Census database.